

Levy model of cancer

Augusto Gonzalez

Instituto de Cibernética, Matemática y Física, Calle E 309, Vedado, La Habana, Cuba

Along an individual lifetime, stem cells replicate and suffer modifications in their DNA content. I model the modifications in the DNA of a single cell as a Levy flight, made up of small amplitude Brownian motions plus rare large-jumps events. The distribution function of mutations has a long tail, in which cancer events are located far away. The probability of cancer in a given tissue is roughly estimated as $aN_{cell}N_{step}$, where N_{cell} is the number of stem cells, and N_{step} – the number of replication steps in the evolution of a single cell. I test this expression against recent data on lifetime cancer risk, N_{cell} and N_{step} in different tissues. The coefficient a takes values between 2×10^{-15} and 2×10^{-11} , depending on the role played by carcinogenic factors and the immune response. The smallest values of a correspond to cancers in which randomness plays the major role.

PACS numbers: 87.19.xj, 05.40.Fb, 87.23.Kg

Spontaneous vs induced mutations. A common knowledge states that both the normal activity of stem cells in a healthy individual, and external agents such as ionizing or ultraviolet radiation, toxic substances (derived from smoking, for example), etc cause mutations [1]. Mutations related to the normal function of the cell are called spontaneous. They are thought to have a random origin, External agents, on their side, are visualized as causes of induced mutations.

The debate about spontaneous (random) and induced mutations and their role in carcinogenesis rose recently [2–10] with the article [2], in which data on lifetime risk of cancer in different tissues, along with the number of stem cells, N_{cell} , and the number of replication steps, N_{step} , are compiled.

The purpose of my paper is to present a model for mutations in stem cells and the genesis of cancer. Emphasis is put on the qualitative aspects. Detailed numerical simulations for different tissues are to be published elsewhere.

The accumulative character of mutations. Authors of Ref. [2] postulate that the probability of a given mutation (they are interested in cancer) should depend on the overall number of replication steps in the tissue. This assumption neglects the history in the evolution of each cell.

In my model, on the contrary, the time evolution of cells defines trajectories, as schematically represented in Fig. 1, where one of these trajectories is drawn as a red dashed line.

The idea about trajectories in the evolution of cells means that there are Markov chains [11] of mutations, where the change in the DNA of a cell at step $i + 1$, x_{i+1} , comes from the change in the previous step plus an additional modification:

$$x_{i+1} = x_i + \delta \quad (1)$$

Measuring changes in the DNA A single strand of human DNA contains around 3×10^9 bases of a four letter alphabet: G, A, T, and C. [1]) In order to measure

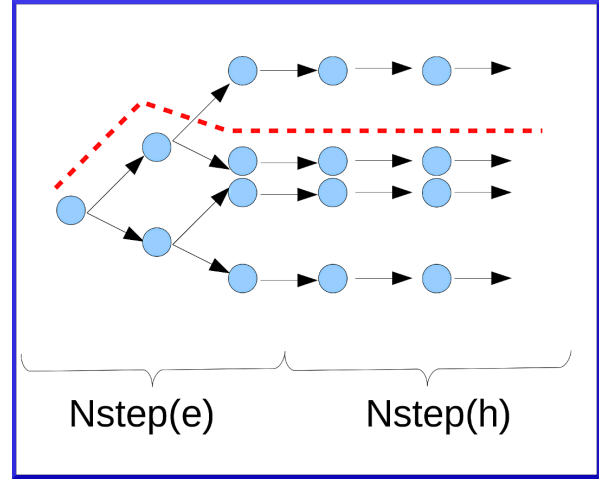


FIG. 1. (Color online) Schematic representation of the evolution of stem cells in a tissue. First, the cells divide until their number reaches N_{cell} . There are $N_{step}^e \approx \log_2 N_{cell}$ steps in this clonal expansion phase. Further on, the number of cells is kept roughly constant. This means that the excess stem cells resulting from divisions go to replace damaged cells in the tissue or to a programmed apoptosis. If there are N_{step}^h steps in this homeostasis phase, then the total number of replication steps along a trajectory is $N_{step} = N_{step}^e + N_{step}^h$.

changes in the DNA, one may use a variable similar to that one of paper [12]).

First, define an auxiliary variable at site α in the molecule: $u_\alpha(G) = 3/8$, $u_\alpha(A) = 1/8$, $u_\alpha(T) = -1/8$, and $u_\alpha(C) = -3/8$. Then, define a walk along the DNA:

$$y(\beta) = \sum_{\alpha=1}^{\beta} u_\alpha. \quad (2)$$

As a function of β , the variable y draws a profile of the DNA molecule, and modifications can be measured as: $X(\beta) = y(\beta) - y_0(\beta)$. where y correspond to the mutated DNA, and y_0 – to the initial configuration. Of

course, there are so many $X(\beta)$, three billions, that they are not of practical use. The strategy could be to restrict the analysis to certain coding regions in the DNA, for example, and for these regions, to use variables measuring global changes or distances to the original function:

$$X = \sum_{\alpha=1}^L (u'_\alpha - u_\alpha), \quad (3)$$

$$X^{(1)} = \sum_{\alpha=1}^L \alpha (u'_\alpha - u_\alpha), \quad (4)$$

$X^{(2)}$ (the second moment), etc. L is the length of the coding region. The Shannon informational entropy [13]) could also be of use.

In what follows, I shall assume that mutations in a given coding region are well characterized by a few global variables.

Other heritable gene variations, not involving changes in the DNA sequence [14], could, in principle, be incorporated, although presently I do not have a proposal for a variable measuring them.

Modeling mutations The δ term in Eq. (1) represents mutations at step $i + 1$. It may come from a partially repaired damage in the DNA that is fixed after replication, or from an undesired error in the replication process. It should be stressed that both the repair mechanisms and the replication process guarantee very high fidelities. The error introduced by the latter, for example, is around one mistaken base per 10^9 bases in the DNA strand [1]).

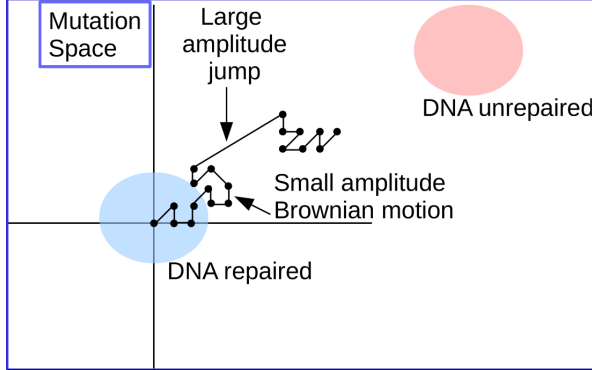


FIG. 2. (Color online) Schematic representation of a single cell mutation trajectory. The starting point is $X = 0$. In the mutation space, I distinguished regions in which the DNA repair mechanism is active or damaged. The latter is one of the hallmarks of cancer, known as genetic instability [15].

Let me stress, once again, that δ is not the damage caused by endogenous or external factors, but the resulting modification after the action of the repair mechanisms, and fixation. It is known, for example, that ionizing radiation may cause double strand breaks in the

DNA [16]. These damages are very difficult to repair [1]). The repair mechanism itself may introduce large changes in the resulting DNA composition after a double strand break event.

My proposal for δ is the following: $\delta = \delta_B + \delta_{LJ}$. The δ_B component corresponds to a Brownian motion with maximal amplitude D_B . Notice that $D_B = 1$ would mean roughly a change of basis in each replication step because $u_\alpha(G) - u_\alpha(C) = 3/4$. This Brownian motion introduces local modifications in the DNA. After N_{step} replication steps, the characteristic dispersion of a trajectory due to this Brownian motion (something like the radius of the colored region near the origin in Fig. 2) is $D_B \sqrt{N_{step}}$. [17]

The large-jump component of δ , δ_{LJ} , on the other hand, is modeled with the help of rare events with total probability $p \ll 1$, and a probability density proportional to $1/\delta_{LJ}^2$, where the amplitude ranges from D_B to infinity (in practice, I will introduce a cutoff, D_{max}). The combination of the Brownian motion and the large amplitude jumps leads to Levy flights [18]) in the mutation space, schematically represented in Fig. 2.

Let me notice that the distribution function associated to Levy flights is a fat- or long-tail one. This fact could be related to the long range correlations observed in the walks along the DNA [12].

The long-tail distribution function of mutations. Four parameters enter my oversimplified Levy model of mutations in stem cells: N_{cell} , N_{step} , D_B and p . In order to fix ideas, I show a calculation with parameters that, although arbitrary, approximately fit the data on lifetime risk of cancer in the gallbladder tissue. The result, however, will be not only the probability of cancer, but the distribution function of mutations of any amplitude.

I took $N_{step}^e = 20$, which corresponds to $N_{cell} \approx 10^6$. On the other hand, $N_{step}^h = 47$, as in Ref. [2], thus $N_{step} = 67$.

I restrict the analysis to a coding region in the DNA of length $L = 10^6$. Around 10^{-3} basis are changed in each replication step, thus I take $D_B = 10^{-3}$. Cancer events are assumed to be at a distance $X_{cancer} \gg D_B \sqrt{N_{step}}$ from the origin. It is fixed to $X_{cancer} = 1000$. The probability $p = 3.8 \times 10^{-5}$ was chosen in order to fit the lifetime risk of gallbladder cancer [2, 19].

Simulations start from a single cell. After N_{step}^e division steps, the N_{cell} trajectories are generated. These trajectories proceed N_{step}^h steps further. In any replication step, either of the expansion or homeostasis phase, mutations are given by Eq. (1), where δ contains both the Brownian and the large-amplitude components.

The probability distribution function for mutations in a cell, $P(X)$, is the probability that a cell arrives at the end point with an amplitude X . I compute not $P(X)$, but the cumulative probability distribution, $P(|X| > Z)$, which is shown in Fig. 3.

The Brownian radius, $\sqrt{N_{step}} D_B \sim 8 \times 10^{-3}$, concentrating most of the points, is apparent in the figure. In

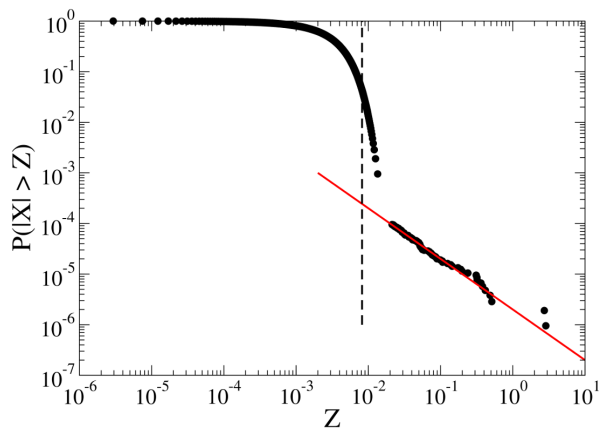


FIG. 3. (Color online) The average cumulative probability of mutations, $P(|X| > Z)$, in a coding segment of the DNA molecule. Points come from the numerical simulations, whereas the red solid line is a $1/Z$ fit to the tail. The Brownian radius, $D_B \sqrt{N_{step}}$, is marked by a dashed line. Parameters are chosen in order that the slope in the tail reproduces the lifetime risk for cancer in the gallbladder tissue.

addition, the tail can be fitted to a $1/Z$ dependence. The coefficient is roughly $N_{step} D_B p$.

Notice that the probability of cancer in the tissue can be estimated as $N_{cell} P(|X| > X_{cancer})$. As X_{cancer} is in the tail of the distribution, we may use the asymptotic formula:

$$risk \approx N_{cell} N_{step} D_B p / X_{cancer} = a N_{cell} N_{step}. \quad (5)$$

For the gallbladder example, $a \approx 2.5 \times 10^{-11}$. Because $a \sim D_B p / X_{cancer}$, there is an arbitrariness in the selection of the parameters D_B , p and X_{cancer} . However, it should be possible to select or pick up a unique meaningful set consistent with this numerical value. The asymptotic formula, Eq. (5), does not depend on the choice of parameters.

Analysis of the data on cancer in different tissues. I use Eq. (5) in order to re-examine the data presented in paper [2]). The qualitative idea is the following: the lowest values of a correspond to tissues in which mutations are close to the naturally occurring ones in a healthy individual. On the contrary, high values of a indicate the presence of strong abnormal conditions or external factors in the genesis of cancer. In Fig. 4, I show the results for the lifetime risk of cancer per stem cell versus N_{step} . This magnitude can be directly related to the formulae of the previous sections. In order to facilitate the analysis, the studied cancers are divided in groups.

Group I includes 11 points in the figure, located in a band delimited by coefficients $2 \times 10^{-14} < a < 10^{-13}$. In the lack of a better name, I call this set the normal group. I use a solid red line in order to distinguish the bottom edge, and a dashed red line for the top one. In

this set, randomness seems to play an important role in the genesis of cancer, as originally claimed in Ref. [2]. The fact that this group is composed by very different tissues – from the medulloblastoma to the colorectal adenocarcinoma – is, perhaps, the confirmation that, under unperturbed conditions, tissues in the body evolve in a very similar way. The starting point in the Levy model, D_B , p and X_{cancer} should take very similar values for all of them.

For the coefficient a , we may write the expression:

$$a = ERS \times 2 \times 10^{-14}. \quad (6)$$

The in-front factor, ERS , a kind of measure of the effects of lifestyle or external carcinogens, takes values between 1 and 5 (see Table I), meaning that, in principle, by means of proper correctives, the risk for cancer could be reduced, for example, around twice in the colorectal adenocarcinoma, four times in the basal cell carcinoma, or five times in the lung adenocarcinoma in the non-smoker sub-population.

Group II, with five points in the figure, include cases in which genetic or viral causes are predominant. Genetic predisposition means that mutations start at a point closer than usual to the cancer region. Thus, the distance X_{cancer} is much shorter and the probability dramatically increases. The ERS index exhibits very high values in this set.

The abnormal values of ERS for the four cases contained in Group III have, in my opinion, an immunological origin. Indeed, germinal cells and the brain are partially isolated from the immune agents. Our body uses barriers in order to protect these tissues against infections, but the barriers can not protect against tumors, which come from inside. From the point of view of cancer, they are immunodepressed tissues.

Thus, I may say that $a_{prot} \approx 2 \times 10^{-14}$ is a reference value for a tissue protected by a normal immune system, whereas $a_{depr} \approx 2 \times 10^{-12}$ (100 times higher) refers to immunodepression conditions.

On the other hand, the extremely low value of a for the small intestine adenocarcinoma (eight times lower than normal) can not have other explanation than overprotection by the immune system. One may speculate that the small intestine is a possible entrance door for the microbiota living in the colon, and as such it requires special protection. Paneth cells, Peyer's patches, and other structures concentrated in the distal ileum, are perhaps the responsables for this reinforced protection. This fact should be further studied. If confirmed, one can even imagine therapies against cancer or other illness exploiting this extra capacity of the small intestine.

Finally, there is a group of 11 cancers (5 tissues) exhibiting abnormally high values of the ERS index, presumably related to strong external factors. One example is lung adenocarcinoma, for which the concurrence of radioactive Radon and smoking produces a 90-fold increase of the slope. The extreme case in this group is

| Cancer type | ERS |
|------------------------------------|---------|
| Group I. Normal | |
| Hepatocellular C | 1.13 |
| Melanoma | 1.16 |
| Pancreatic endocrine C | 1.23 |
| Pancreatic ductal AC | 1.45 |
| Medulloblastoma | 1.49 |
| Myeloid leukemia | 1.54 |
| Duodenal AC | 1.93 |
| Lymphocytic leukemia | 1.95 |
| Colorectal AC | 2.04 |
| Basal Cell C | 4.02 |
| Lung AC (non-smokers) | 5.15 |
| Group II. Viral and Genetic | |
| Hepatocellular C with HCV | 11.29 |
| Colorectal AC with Lynch | 21.30 |
| Head and Neck SCC with HPV | 122.96 |
| Colorectal AC with FAP | 204.51 |
| Duodenal AC with FAP | 225.29 |
| Group III. Immune | |
| Small intestinal AC | 0.12 |
| Glioblastoma | 14.48 |
| Testicular germinal cell | 52.78 |
| Ovarian germinal cell | 79.86 |
| Group IV. Abnormal | |
| Head and Neck SCC | 21.38 |
| Osteosarcoma (Head) | 70.02 |
| Esophageal SCC | 79.44 |
| Thyroid medullary C | 84.22 |
| Lung AC (smokers) | 92.77 |
| Osteosarcoma (Arms) | 124.72 |
| Osteosarcoma (Pelvis) | 138.08 |
| Osteosarcomas | 153.04 |
| Thyroid papillary and follicular C | 239.78 |
| Osteosarcoma (Legs) | 266.49 |
| Gallbladder non papillary AC | 1299.58 |

TABLE I. The Extra Risk Score (ERS) index of Eq. (6) for cancer in different tissues.

Concluding remarks. In my model, stem cells draw Levy flights in the mutation space. The small amplitude Brownian component is characterized by a radius $D_B \sqrt{N_{step}}$, whereas the rare large jumps give rise to a long tail $\sim 1/Z$ in the cumulative probability distribution. Cancer events are located in the tail. Their rate can be estimated from Eq. (5), where $a \sim D_B p/X_{cancer}$. Variations in a are mainly related to variations in p , the probability of nonlocal changes in the DNA. Trajectories in the mutation space are always random, external carcinogens basically increase the probability p . The ERS index defined in Eq. (6) has, thus, a clear meaning as a reference to a normal tissue, unperturbed by external factors. Re-examination of the data reported in Ref. [2] re-

gallbladder non-papillary adenocarcinoma, with an index $ERS = 1300$, the understanding of which is a real challenge.

Mutations in bacteria. With appropriate parameters, my Levy model can be applied as well to mutations in bacteria. I recall the extremely interesting Long Time Evolution Experiment with *E. Coli*, conducted by Prof. R. Lenski and his group [20]. Besides many other results, they report frequencies at which a mutation with damages in the DNA repair mechanisms becomes dominant in a population [21]. This mutator phenotype has in common with cancer, besides DNA instability, that it should be far away in the tail of the probability distribution. Thus, I may use Eq. (5) for the probability of occurrence, and determine the coefficient a .

The number of cultures they use is small, 12. Thus, I expect statistical errors of the order of $1/\sqrt{12} \sim 0.3$ for the probability. Nevertheless, they report that the mutator phenotype becomes dominant in two cultures (cumulative probability 1/6) when $N_{step} \approx 2500 - 3000$, in a third culture (cumulative probability 1/4) when $N_{step} \approx 8500$, and in a fourth culture (cumulative probability 1/3) when $N_{step} \approx 15000$. From this data and the number of evolving trajectories, $N_{cell} \approx 5 \times 10^6$, I obtain $a_{bact} \approx 5 \times 10^{-12}$. It is remarkable, that a_{bact} is of the same order of magnitude of a_{depr} . Details can be found in Ref. [22].

veals groups of cancers, which range from normal tissues (randomness dominated cancers) to abnormal tissues, in which external carcinogens play the major role. Particularly interesting is the small intestine, which seems to be overprotected by the immune system. My model stresses the role of mutations in the genesis of cancer. It is reasonable to expect, however, a formula like Eq. (5) to be valid even when epigenetic or microenvironment factors are taken into account [23].

Acknowledgments. The author is grateful to A. Cabo and E. Heiden for useful discussions. Support from the National Program of Basic Sciences in Cuba, and from the Office of External Activities of the International Centre for Theoretical Physics (ICTP) is acknowledged.

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell* (Garland

Science, New York, 2002).

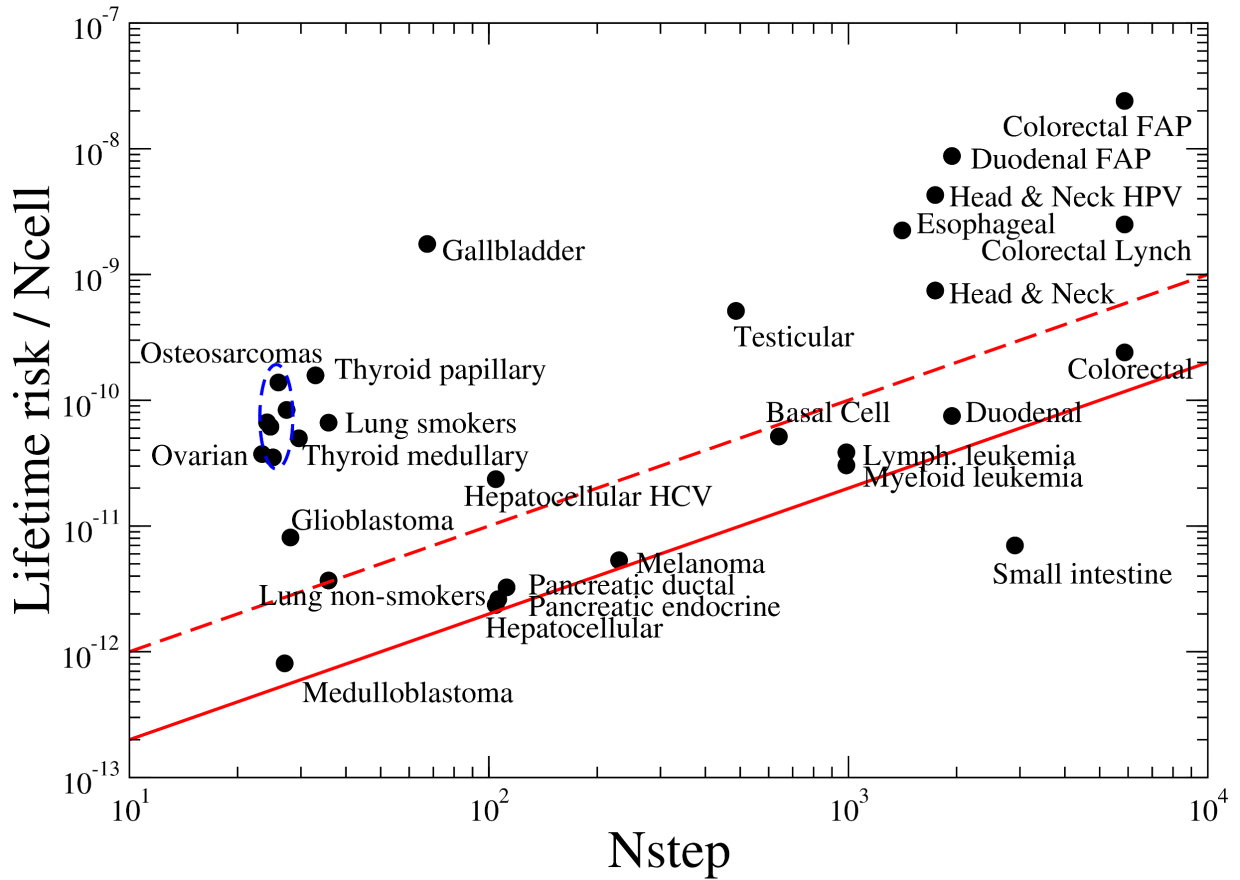


FIG. 4. (Color online) Lifetime risk of cancer per stem cell in a tissue vs N_{step} . The analysis is based on Eq. (5). See the explanation in the main text.

- [2] C. Tomasetti and B. Vogelstein, *Science* **347**, 78 - 81 (2015); Supplementary materials: www.sciencemag.org/content/347/6217/78/suppl/.
- [3] John D. Potter and Ross L. Prentice, *Science* **347**, 727 (2015).
- [4] Nicholas A. Ashford, Patricia Bauman, Halina S. Brown, et al, *Science* **347**, 727 (2015).
- [5] Christopher Wild, Paul Brennan, Martyn Plummer, et al, *Science* **347**, 728 (2015).
- [6] Carolyn Gotay, Trevor Dummer, and John Spinelli, *Science* **347**, 728 (2015).
- [7] Mingyang Song and Edward L. Giovannucci, *Science* **347**, 728-729 (2015).
- [8] Michael O'Callaghan, *Science* **347**, 729 (2015).
- [9] Cristian Tomasetti and Bert Vogelstein, *Science* **347**, 729-731 (2015).
- [10] C. Tomasetti and B. Vogelstein, arXiv:1501.05035.
- [11] V.S. Koroliuk, N.I. Portenko, A.V. Skorojod, and A.F. Turbin, *Handbook on probability theory and mathematical statistics* (Nauka, Moscow, 1978).
- [12] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, et. al., *Physica A* **191**, 25 - 29 (1992).
- [13] T. D. Schneider, *Information and entropy of patterns in genetic switches*, in G. J. Erickson and C. R. Smith, Eds., *Maximum-Entropy and Bayesian Methods in Science and Engineering*, volume 2, pages 147 - 154 (Kluwer Academic, Dordrecht, 1988).
- [14] Adrian Bird, *Nature* **447**, 396-398 (2007).
- [15] D. Hanahan and R.A. Weinberg, *Cell* **100**, 57 - 70 (2000).
- [16] Leon Mullenders, Mike Atkinson, Herwig Paretzke, Laure Sabatier and Simon Bouffler, *Nature Reviews Cancer* **9**, 596 - 604 (2009).
- [17] A. Einstein, *Investigations on the theory of the Brownian movement* (Dover, New York, 1956).
- [18] M.F. Shlesinger, G. Zaslavsky, and U. Frish, Eds., *Levy flights and related phenomena in Physics*, Lecture Notes in Physics, Vol. 450 (Springer, Berlin, 1995).
- [19] Data is taken from <http://seer.cancer.gov/statfacts/>
- [20] R.E. Lenski, *Summary data from the long-term evolution experiment*, <http://myxo.css.msu.edu/ecoli/summdata.html>
- [21] R.E. Lenski, *Phenotypic and genomic evolution during a 20000 generation experiment with the bacterium E. Coli*, in J. Janick, Ed., *Plant Breeding Reviews* **24**, Part 2, pages 225 - 265 (2004).
- [22] A. Gonzalez, *The long-tail distribution function of mutations in bacteria*, in E. Altshuler and J.O. Fossum, Eds., *Proceedings of the Meeting on Complex Matter Systems*, Havana, June 2015. To appear in *Revista Cubana de Fisica*. arXiv:1507.06920.
- [23] Lucio Luzzatto and Pier Paolo Pandolfi, *New England Journal of Medicine* **373**, 84 (2015).